Predicting Health Outcomes Using Machine Learning
Biostatistics 794, Fall 2025

Overview

To consolidate the main skills in this course – framing a health-relevant question, selecting appropriate ML methods, preventing leakage/overfitting, tuning and validating models, interpreting outputs (including variable importance and model behavior), and communicating results responsibly – you will complete an analysis of a real dataset. This project mirrors real collaboration: messy data and constraints, methodological trade-offs, and defensible decisions.

For the project, you will complete an end-to-end machine learning (ML) analysis in **R** on a dataset of your choice. The project has four milestones:

1. a brief topic and data description (due October 3rd),
2. a short research plan (due October 17th),
3. an **8-page** paper (due November 20th), and
4. an in-class presentation (during the final week).

Your primary goal is to design, validate, and interpret a predictive (or unsupervised) ML workflow that answers a clear question.

You may work in teams of two (no three-person teams). The team will choose the project topic.

Choosing a topic (guidance)

- **Pick a subject that interests you**—engagement improves your work and the class discussion.
- **Choose a new problem**, not something you have already done.
- **Data do not have to be public health**; if a topic isn't related, include how the methods used could be used in public health.
- This project aims to **implement the methods from this class**. Adding topics outside scope is usually not a good use of effort (and typically will not improve your grade).
- **Depth over breadth**– whenever you have questions about what to investigate, choose to explore **one question in depth** rather than many associations. It is better to say something *definitive* about one relationship than nothing about three.

Technical expectations (for all projects)

- Use at least one **ML method beyond standard linear/logistic regression** (these may be baselines). Examples: trees (CART), random forests/GBM, SVM, k-NN, naive Bayes, penalized regression (ridge/lasso/elastic net), GAMs, clustering (k-means/hierarchical), PCA/autoencoders, simple neural nets, survival ML (e.g., random survival forests), text embeddings/NLP.
- **Validation:** predeclare splits; use cross-validation (e.g., k-fold/nested) or a hold-out test; report suitable metrics (AUC/PR-AUC, Brier/calibration, RMSE/MAE) aligned with the goal. Avoid tuning on the test set.
- **Leakage checks:** define features without outcome leakage; respect time/order and site boundaries as applicable; handle class imbalance transparently.
- **Interpretability:** use variable importance, partial dependence/accumulated local effects, or SHAP (when appropriate).
- **Ethics & privacy:** do not place PHI/identifiers in public tools; describe de-identification and any fairness checks (if relevant).
- **Reproducibility:** submit R code and a README; set seeds; include `sessionInfo()`; cite packages.
- **Generative AI:** allowed per course policy; *disclose* any AI assistance in your README and ensure you understand and verify all outputs.

Project types (data expectation)

- *Static type:* projects using data that require *straightforward* cleaning or variable transformations; commonly "well-worn" datasets.
- *Dynamic type:* projects where the data require *meaningful* cleaning, transformations, or exploration (e.g., linkage, deduplication, NLP preprocessing) before use. Dynamic projects must **demonstrate** the data background, acquisition, and cleaning decisions in both the paper and the presentation.

If your project is a static type, my expectation of the modeling depth and validation is higher than with a dynamic type. Both types will fit some model to a dataset, but with a static type, I expect more in terms of comparisons to other methods, the number of predictive accuracy metrics considered, interpretations (quantitative and graphical), etc.

Final paper

The final paper will be at most 8 pages, excluding references and appendix. Below is a ***guide*** (not a rulebook):

1. Introduction

Background & motivation; brief literature; **specific aims** (bulleted).

2. Data

Source, population, inclusion/exclusion; variables (outcome(s), predictors, timing); how the analytic sample was obtained; missing-data mechanisms and handling (e.g., complete-case, imputation). *Dynamic projects:* describe acquisition/cleaning pipeline and key choices.

3. Methods

Analysis plan: EDA → feature engineering → splitting/validation → primary model(s) → baselines.
Models: method rationale, key hyperparameters, tuning approach (grid/random/Bayesian, early stopping).
Performance metrics: why they fit the goal (discrimination, calibration, error).
Interpretability plan: e.g., VIP, PDP/ALE, SHAP.
Risk controls: leakage, class imbalance, site/temporal validation; fairness checks if relevant.

4. Results

EDA highlights (concise figures/tables); tuning/selection results; baseline comparisons; hold-out/CV performance with uncertainty (e.g., CIs via CV/bootstrap); calibration and error analysis; interpretability findings (most influential features; stability).

5. Discussion & Conclusion

Answer the aims; practical significance; limitations and threats to validity (measurement, shift, small-n, confounding); transportability/generalization; next steps or deployment considerations; ethical notes (privacy, bias, transparency).

Figures/Tables

Figures and tables must be **meaningful** (not decorative), labeled, and readable; avoid chartjunk.

Oral presentation (final week)

Timing: typical slot is 20-25 minute talk + 3–5 minutes Q&A.
Suggested flow:

1. Problem and why it matters; **why ML** is appropriate here.
2. Data & design (what, who, when; splits; leakage risks).
3. Method (conceptual; what you tuned; how you validated; metrics).
4. Results (performance, calibration, key drivers, error analysis).

5. Limitations, ethics, and one concrete "how I'd improve it" idea.

Style tips: minimize text; include a one-slide pipeline; use clear figures/tables; rehearse timing; be prepared to explain one core figure without notes.

Milestones & submissions

- **Topic description:** (due October 3rd) 1–2 paragraphs (question, data source, outcome, predictors, why ML).
- **Research plan:** (due October 17th) ≤1 page (split/validation plan, primary model & baseline(s), metrics, interpretability & leakage checks, anticipated challenges).
- **Final paper:** (due November 20th) 8 pages + code and README.
- **Slides:** (during the final week) PDF uploaded before class.

Grading rubric (100 points total)

Milestones (20 pts)	
Topic description: clarity & feasibility	10
Research plan: specificity & validity	10
Final paper (55 pts)	
Problem framing, literature, aims	8
Data description & missing-data handling	7
Methods: appropriateness, tuning/validation rigor, leakage controls	15
Results: performance, calibration, interpretability, baselines	15
Discussion: limitations, ethics, transportability, conclusions	10
Presentation (20 pts)	
Organization, clarity, visuals, timing	8
Correctness & explanation of methods/metrics	6
Insightfulness of interpretation & Q&A	6
Reproducibility & professionalism (5 pts)	
Clean, runnable R code + README; seeds/session info; AI-use disclosure	5

Notes:
Static projects are expected to excel in **modeling depth and validation**.
Dynamic projects are expected to document and justify **data engineering choices** (without sacrificing validation quality).
Baselines (e.g., logistic regression, clinical score) are encouraged for context but should not be the main method.

Questions are welcome in office hours or on the course forum. Please follow the course Generative AI policy and disclose any AI assistance.